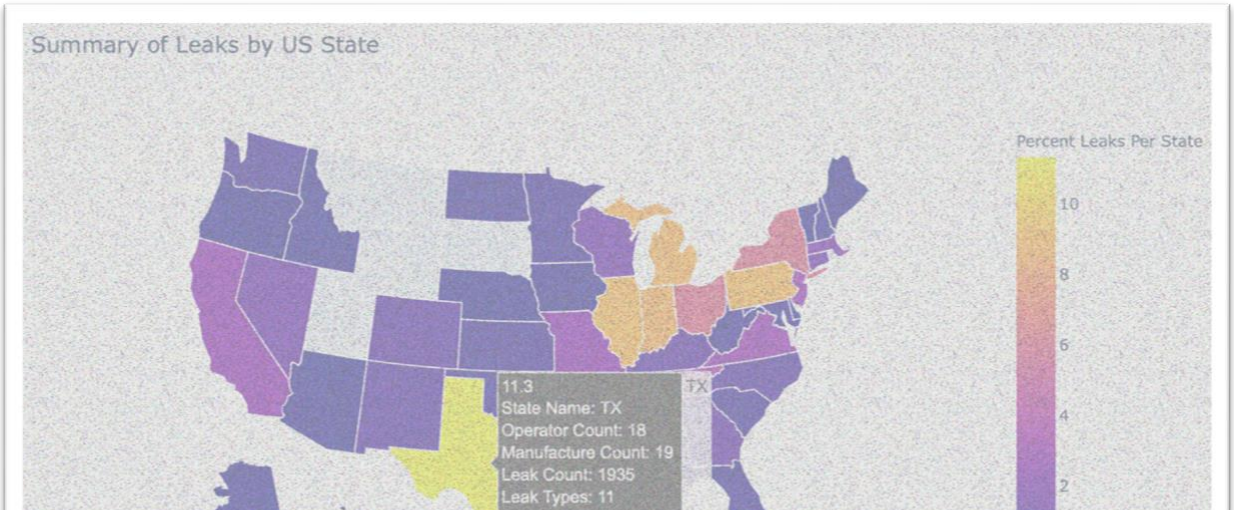
Applying Deep Learning to Detect Leak Cause in Gas Pipelines and Predict Failure Time Frame: Part-2

MECHANICAL FITTING FAILURE CLASSIFICATION PROBLEM

Benefits:

- Classified multi-leak cause with 90.18% accuracy.
- Predicted average life of mechanical fittings at 41.75 years.
- Texas and Washington DC recorded 22.6% most leaks in US.



Prepared by:
Prashant Sanghal

Data Preparation and Model Performance:

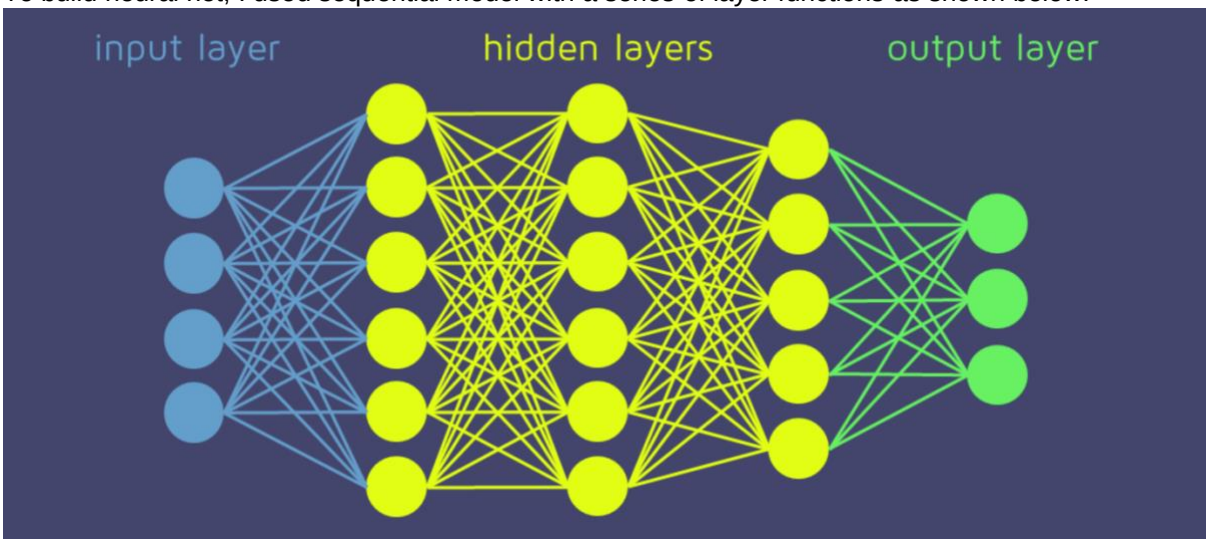
In connection with future recommendations discussed in part-1, the first step in the process, was to get the data ready to try out deep learning model. Since, our dataset contains multiple categorical values, I built two separate models using similar approach I tried earlier. First model, I tried it with label encoder and one-hot encoding while the second model I used binary encoder, which reduced the number of columns from 3,910 to 140 as well as training run time from 10 minutes to under 10 seconds. Moreover, the model performance using binary encoder improved classification accuracy from 46.1% to 90.18% and predicted failure timeframe with MSLE score of 0.26 on the validation data set.

MSLE, which is a loss function 'mean_squared_logarithmic_error', measured the difference between actual and predicted time frame failure range. Our goal here was to fine tune the model so that we could optimize our weights and biases at local minima.

Building Deep Learning Model:

- **Architecture:**

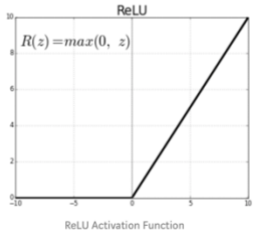
To build neural net, I used sequential model with a series of layer functions as shown below:



First, introduced input layer with fixed numbers of neurons and input shape retrieved from 'mechanical_fitting_file_cleaned.csv' dataset prepared earlier.
Second, added hidden layer which computed weighted inputs for each neuron and used activation function "relu" to generate final input weights for the output layer.
Third, this is the final output layer where I used activation function 'softmax' to classify multi-leak labels and 'relu' to predict failure timeframe of the leak.

$$g(z) = \max\{0, z\}$$

Activation functions 'softmax' and relu (rectified linear activation function) effectively deals with multi-label classification and vanishing gradient problems in predictive models.



ReLU Activation Function

Finally, to avoid model overfitting, between each layer, I used 'dropout' technique to randomly ignore 2% to 3% of the neurons in the network, meaning while training, some neurons were not weighted or updated during forward and backward propagation cycle. This is important feature of neural net which forced other neurons in the network to step-in for the missing neurons, thereby, building a better generalized model non-sensitive to specific weights or overfitting.

Another important consideration was selecting the number of neurons or nodes in each layers as discussed above. As a thumb rule, I used average of features and labels as a starting point and ran few iterations to optimize neuron units.

- **Model Compile:**

  Once, neural net architecture was complete, I compiled the model using ‘adam’ optimizer, loss function ‘cross entropy’ and metrics ‘accuracy’ to classify leak cause labels While, for our failure time frame prediction problem, the loss-function had to be changed from ‘cross entropy to 'mean_squared_logarithmic_error' and ‘accuracy’ matrices had to be removed.

- **Fit model:**

  After compilation, the model was ready to be trained. I split the data in to 80% training and 20% validation data set and experimented with different values of batch_sizes, epochs to evaluate model performance. I also used ‘early stopping’ callbacks to cut training run time.

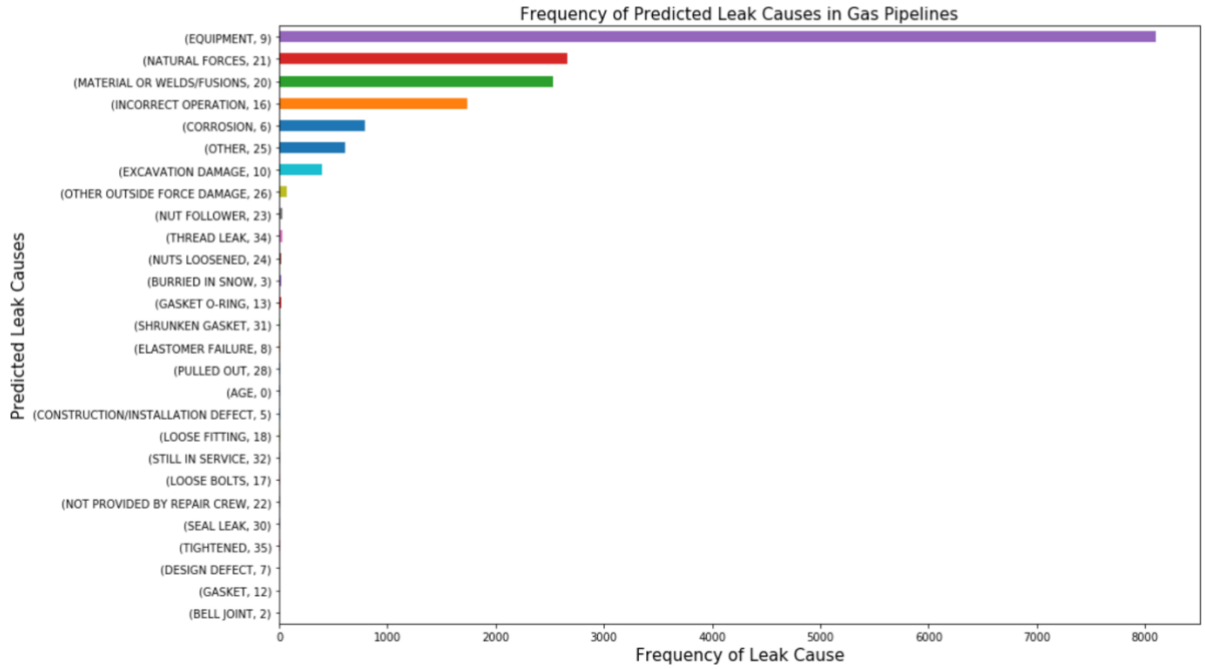
Model Prediction:

After training the model, we were able to classify leak causes and predict failure timeframe across multiple US locations on the map. The main focus was to see if we could summarize model findings visually.

A. Key takeaways: Classifying Leak Cause

- **Frequency of Predicted Leak Cause:** Label 9 (Equipment), 21 (natural Forces) and 20 (Welding) are top 3 reasons for leaks in gas pipeline, as seen earlier during EDA.
- **Neural Net Accuracy:** 90.18% on validation dataset.

```
predicted_leak_cause_text  predicted_labels
EQUIPMENT                  9                  8102
NATURAL FORCES             21                 2663
MATERIAL OR WELDS/FUSIONS  20                 2529
INCORRECT OPERATION        16                 1742
CORROSION                  6                   795
Name: predicted_labels, dtype: int64
```

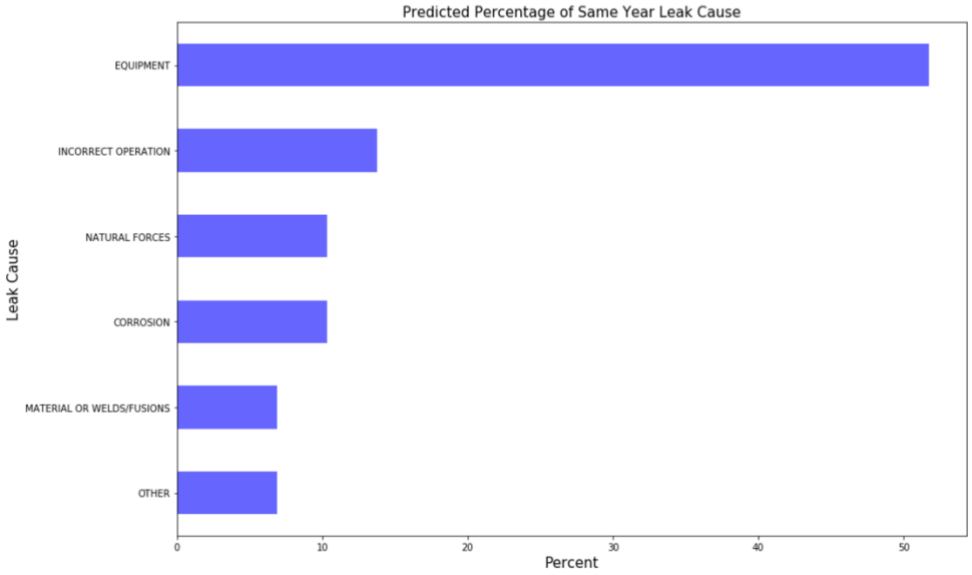


B. Key takeaways: Predicting Failure Timeframe:

- **Same Year Leak Causes:** Classified 20 counts of equipment failures within the same year.

Same Year Leak Casue:

	count	percent
EQUIPMENT	15	51.72
INCORRECT OPERATION	4	13.79
CORROSION	3	10.34
NATURAL FORCES	3	10.34
OTHER	2	6.90
MATERIAL OR WELDS/FUSIONS	2	6.90

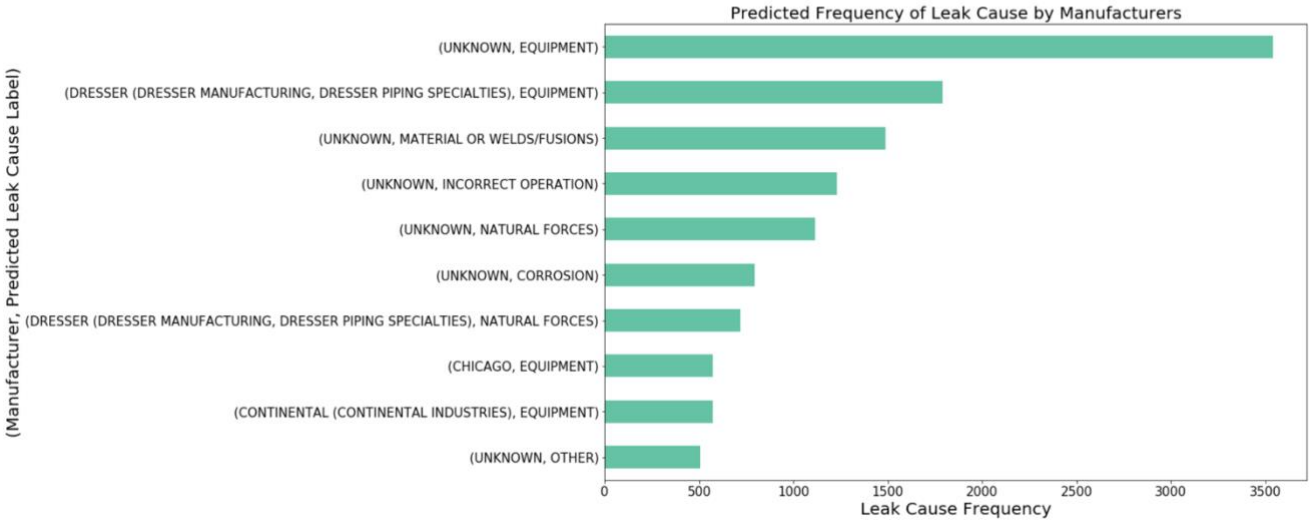


- **Frequency of Leak Cause by Manufacturer:** Dresser Equipment showed 1,791 failed cases.

Predicted Frequency of Leak Cause by Manufacturers:

MANUFACTURE	predicted_leak_cause_text	
UNKNOWN	EQUIPMENT	3543
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	EQUIPMENT	1791
UNKNOWN	MATERIAL OR WELDS/FUSIONS	1490
	INCORRECT OPERATION	1231
	NATURAL FORCES	1117

Name: predicted_leak_cause_text, dtype: int64

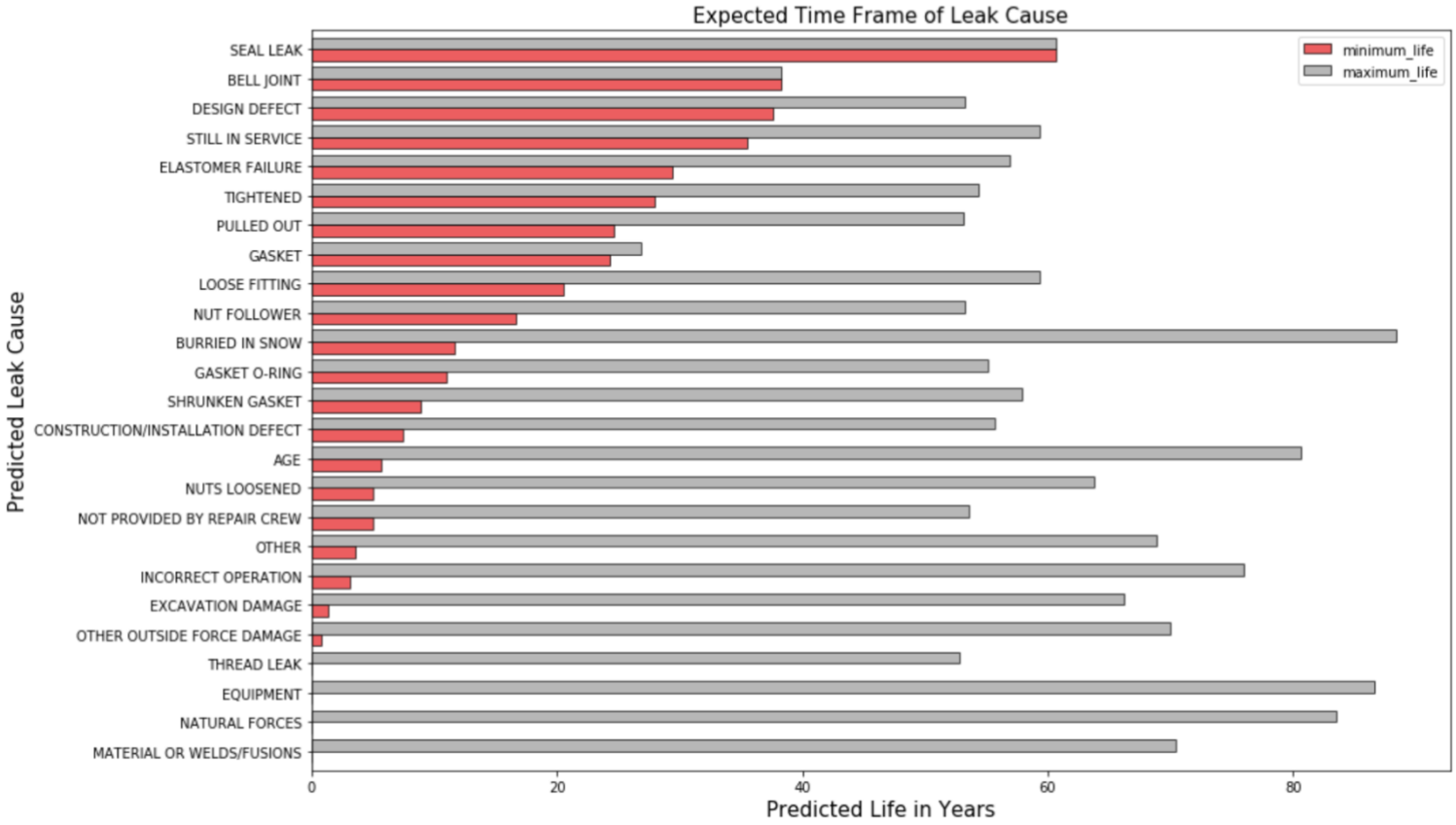


- **Expected Time Frame of Leak Cause:** Average age of mechanical fitting range between 15.18 to 61.8 years with an average life of 41.75 years. Leaks caused due to thread leak, equipment, natural forces and welding failures have high life variations. It could fail within the same year or last an average life of 61.8 years depending upon severity of defect, environmental conditions and time of discovery.

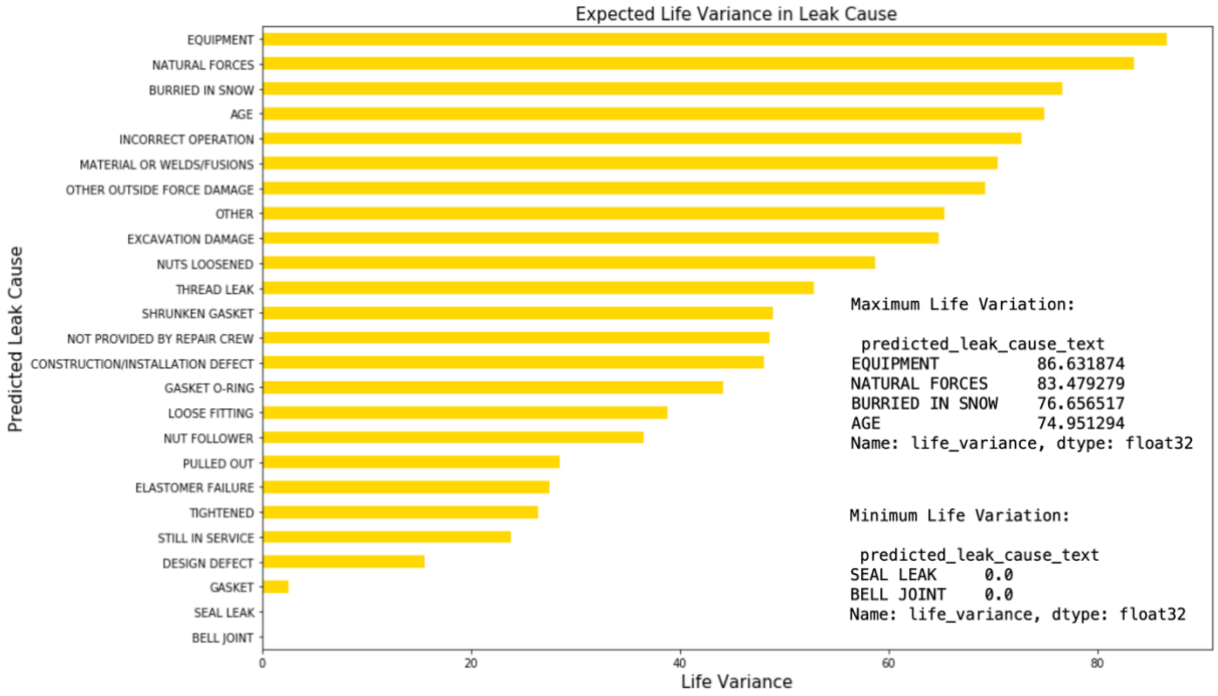
Table showing predicted Time Frame of Leak Cause:

predicted_leak_cause_text	minimum_life	maximum_life
BURRIED IN SNOW	11.713003	88.369522
EQUIPMENT	0.000000	86.631874
NATURAL FORCES	0.000000	83.479279
AGE	5.653286	80.604584
INCORRECT OPERATION	3.153261	75.921120

Maximum Average Life of Mechanical Fittings: 61.8
Minimum Average Life of Mechanical Fitting: 15.18
Average Predicted Life of Fittings 41.75

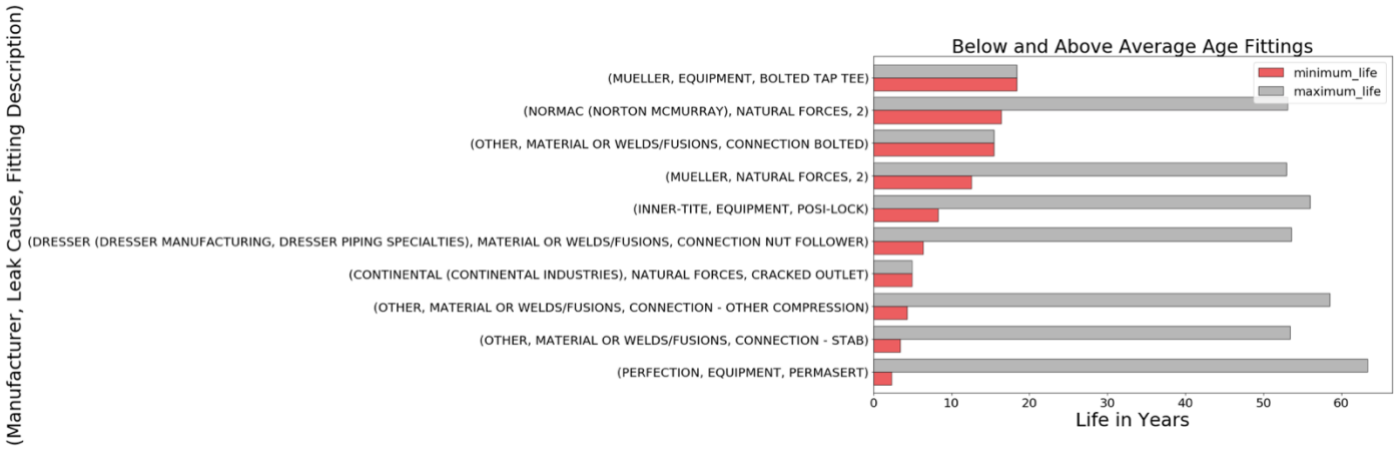


- **Expected Life Variation in Leak Cause**: Leaks caused due to equipment, natural forces and buried in snow showed maximum life variation while seal leak and bell joint zero variation with an average life of 49.50 years. We need to study the reason for high life variance and why some fittings failed before the minimum average life of 15.18 years.



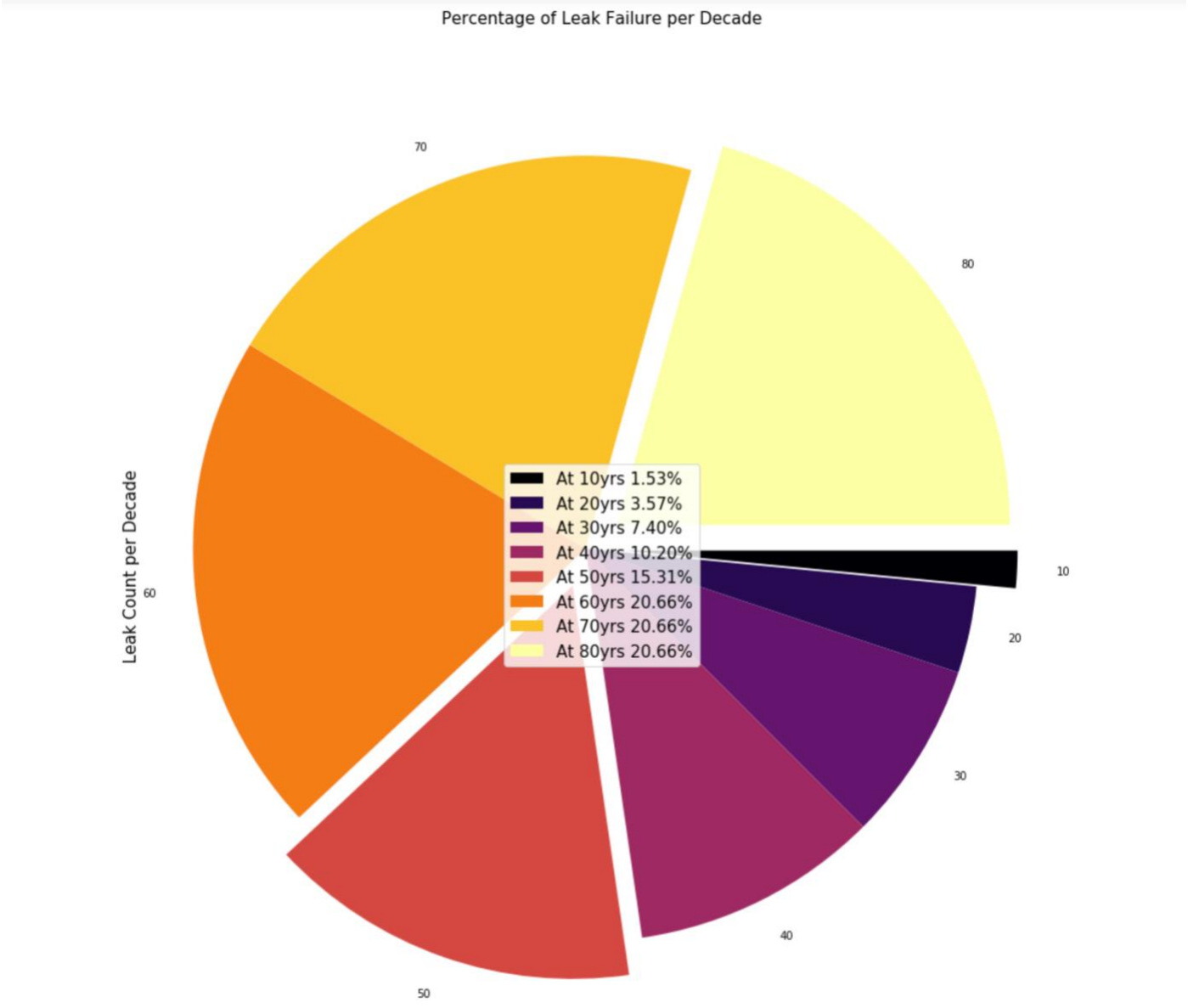
- **Short & Long-Lived Fittings**: Fittings which failed before minimum average life of 15.18 years and lived longer than average age of 41.75 years.

			minimum_life	maximum_life	supply_count
MANUFACTURE	**predicted_leak_cause_text**	**LOT_ATTRIBUTES**			
DRESSER (DRESSER MANUFACTURING, DRESSER PIPING SPECIALTIES)	**MATERIAL OR WELDS/FUSIONS**	**CONNECTION NUT FOLLOWER**	6.412117	53.592041	6
INNER-TITE	**EQUIPMENT**	**POSI-LOCK**	8.305901	55.978275	30
MUELLER	**NATURAL FORCES**	**2**	12.569158	52.918533	5
OTHER	**MATERIAL OR WELDS/FUSIONS**	**CONNECTION - OTHER COMPRESSION**	4.319154	58.520718	9
		CONNECTION - STAB	3.412651	53.448997	5
PERFECTION	**EQUIPMENT**	**PERMASERT**	2.322948	63.355377	26



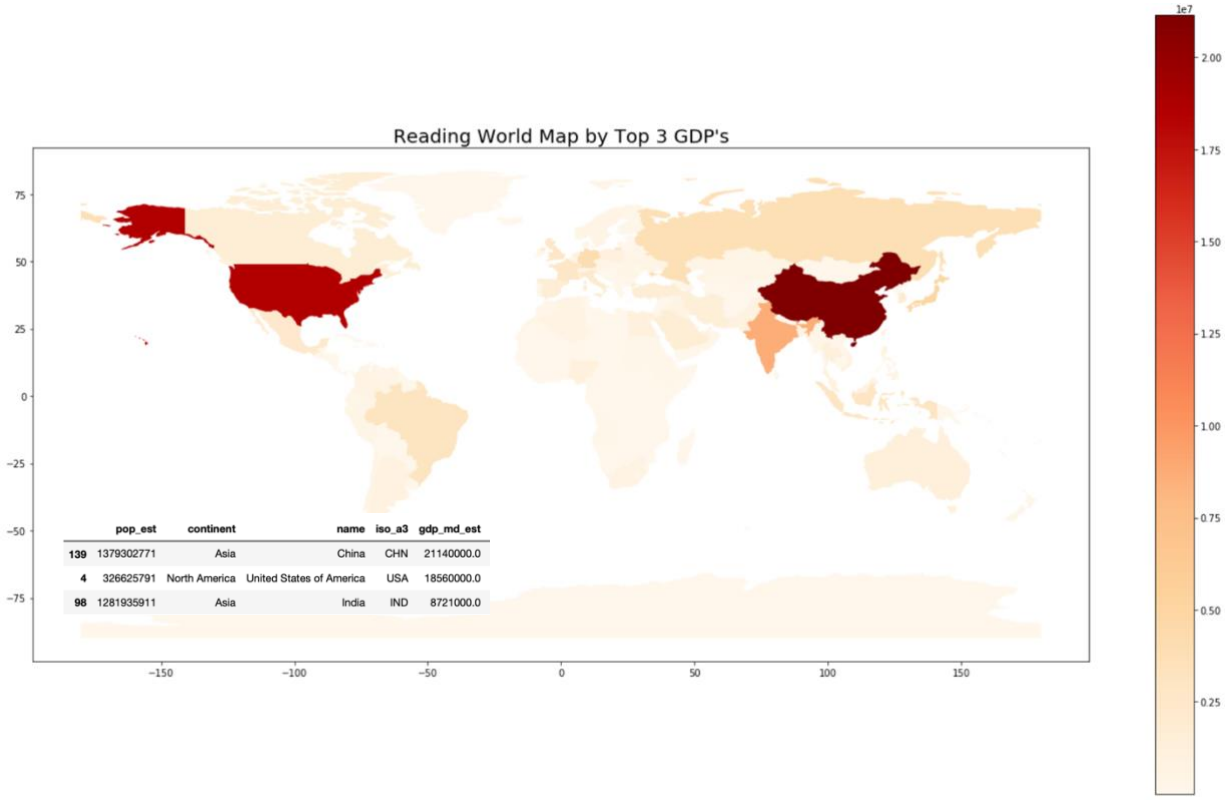
- **Percent of Leaks Every 10 Years:** At 10 years 1.53% and at 80 years 20.66% pipelines could leak due to mechanical fitting failures.

at_life_yrs	total_leak_count	percent_leak_decade
10	6	1.53
20	14	3.57
30	29	7.40
40	40	10.20
50	60	15.31
60	81	20.66
70	81	20.66
80	81	20.66



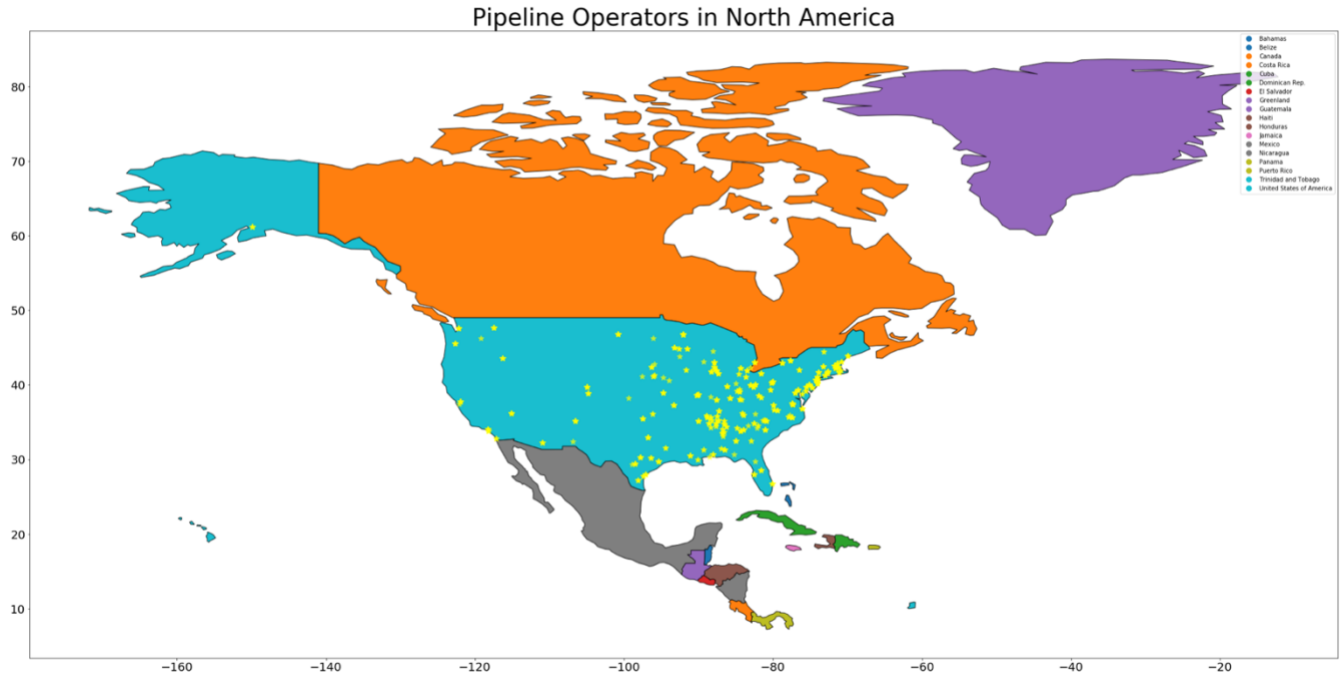
C. Key Takeaways: Geographical Locations

- **Top 3 GDPs:** China, USA and India are the top 3 GDP's in the world. This report will summarize gas pipelines data for north american region.

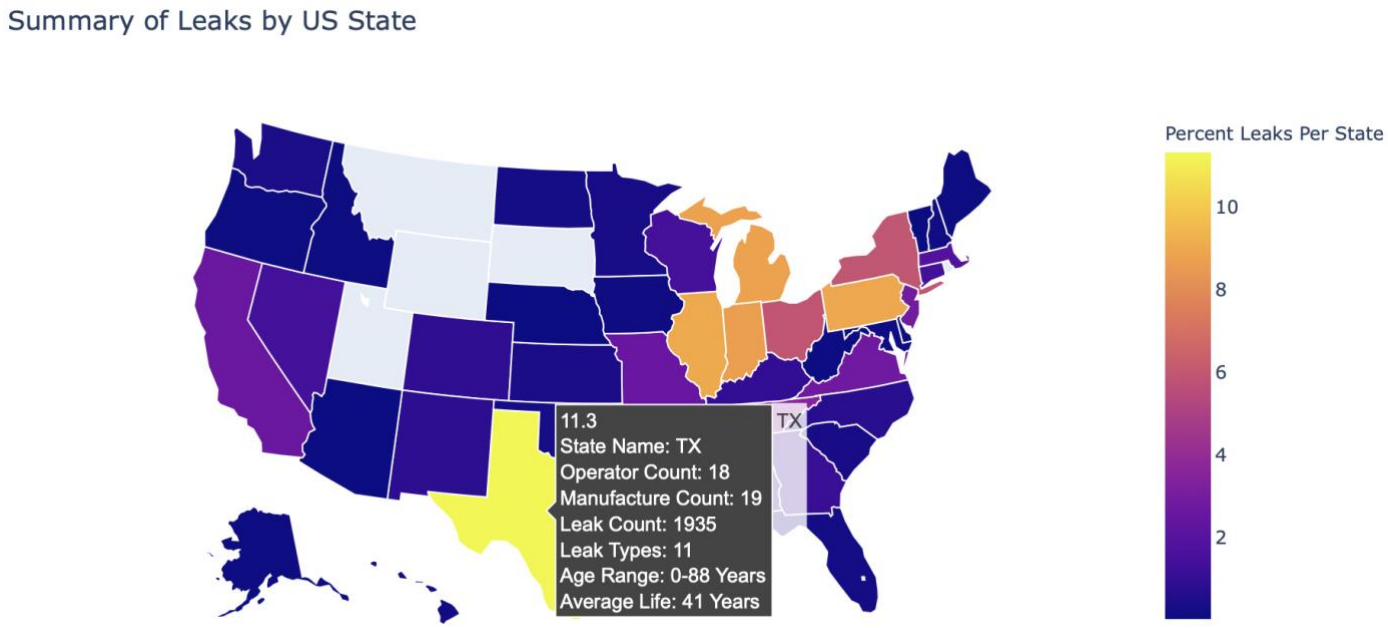


	pop_est	continent	name	iso_a3	gdp_md_est
139	1379302771	Asia	China	CHN	21140000.0
4	326625791	North America	United States of America	USA	18560000.0
98	1281935911	Asia	India	IND	8721000.0

- **Pipeline Operators in North America:** Showing pipeline location coordinates **in** USA and Alaska represented by yellow dots.



- **Summary of Gas Pipeline Leaks Predicted by Neural Net Across US:** Texas and Washington DC recorded maximum leaks at 11.3% and 11.34% with average life of 41 years while Vermont represented lowest leak region at 0.01% with minimum life of 51 years.



Conclusion:

- Using deep learning model, classified multi-leak cause with 90.18% accuracy and predicted failure timeframe with MSLE score of 0.26.
- Predicted 41.75 years as average life of mechanical fittings, which can be used as a baseline to study/optimize age variations found in the fittings.
- Finally, mapped geographical locations and summarized predictive insights to quickly identify gas leak regions.

---End of Project Report---